CHAPTER 5

RANDOM SAMPLING AND SAMPLING DISTRIBUTIONS

TABLE OF CONTENTS

Selection Bias and Sampling

We now return to the primary objective of statistics -- namely, the use of sample information to infer the nature of a population. Predicting your company's annual sales from its sales records of the last 10 years, the annual return on an investment from the results of a number of similar investments, and an evening's winnings at blackjack based on your previous (successful and unsuccessful) treks to the casino are examples of statistical inferences, and each involves an element of uncertainty.

Defining a Common Language about Sampling

- The **population** (or *universe*) is the entire collection of *units*, (individuals or objects or the list of measurements) about which we would like information.
- The **sample** is the collection of units we will actually measure or the collection of measurements we will actually obtain.
- The **sampling frame** is a list of units from which the sample is chosen. Ideally, it includes the whole population.
- In a **sample survey**, measurements are taken on a sample from the population.
- A **census** is a survey in which the entire population is measured.

A **parameter** is a numerical descriptive measure of a population. A **sample statistic** is a numerical descriptive measure of a sample. Since it is almost always too costly and/or time-consuming to conduct a census of a population and compute its parameters, we use the information contained in a sample to make inferences about the parameters of a population.

How a *sample* is selected from a *population* is of vital importance. Since statistics vary from sample to sample, any inferences base on them will necessarily be subject to some uncertainty. How, then, do we judge the reliability of a sample statistic as a tool in making an inference about the corresponding population parameter?

First, some basic notions of sampling. We'd like the sample to be **representative** of the population. If we select n elements as a sample from a population, how can we tell whether the sample is representative? Sometimes it is hard to judge whether a sample is representative. What other criteria could be used to select valid samples from a population of interest?

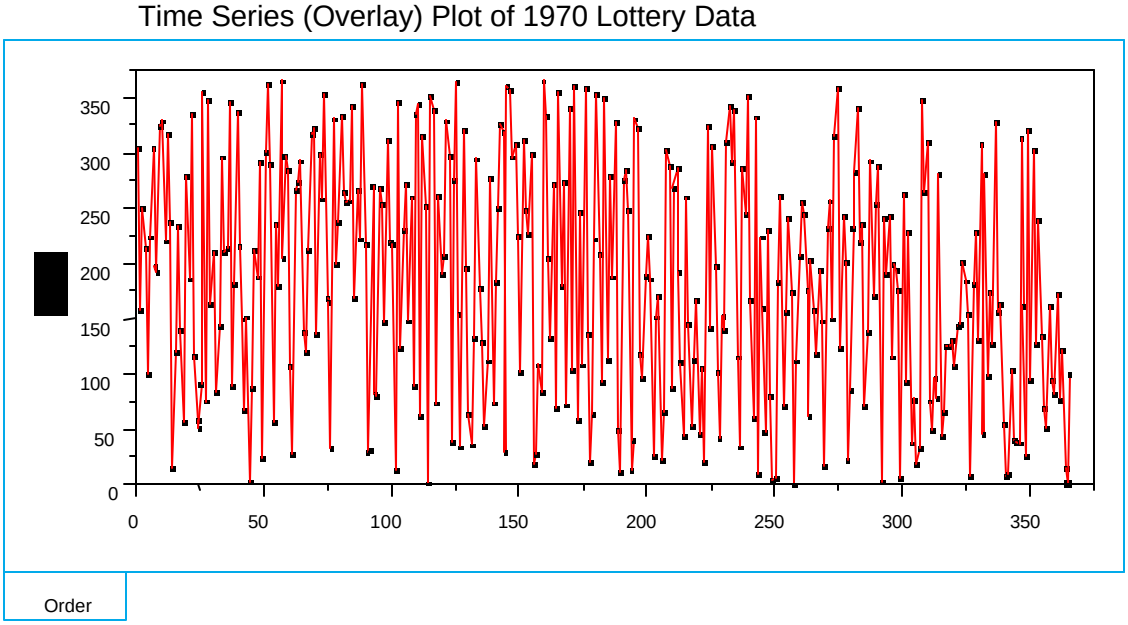
One of the simplest and most frequently used sampling procedures produces what is known as a **random sample**. Suppose we take a sample of n elements from a population. If some samples are more likely to be selected than others, then the sampling is not random. If our method of sampling ensures that every possible combination of n elements in the population has an equal chance of being selected, the n elements are a random sample. Strictly, speaking, this is a **simple random sample**.

Statistics Pre-Term Program – Chapter Five

Random Sampling and Sampling Distributions

Example 5.1. 1970 Draft Lottery[1] [LOTTERY data file]

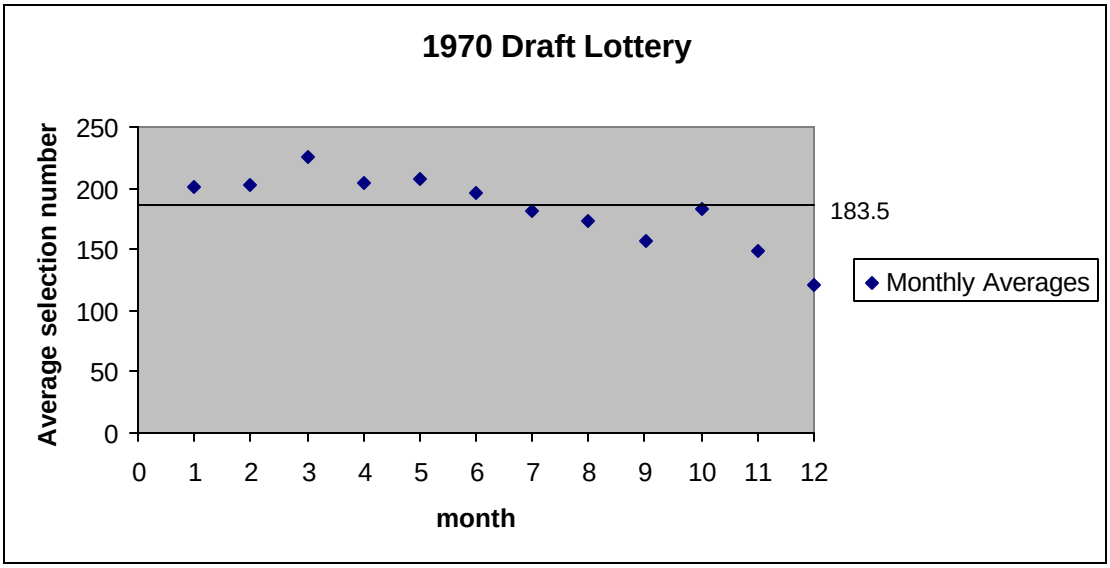
From 1948 through about 1969, men were drafted by age -- oldest first, starting with 25-year olds. On December 1, 1969, the Selective Service System conducted a lottery to determine the order of selection for 1970. 366 possible days in a year were written on slips of paper and placed in egg-shaped capsules. The capsules were then drawn one by one to obtain the order of induction. So the lottery assigned a rank to each of the 366 birthdays. Does this seem like a fair method of determining the order of selection? Below is the summary of the data for the resulting order. Does it appear to provide a fair order of selection?



Month	Mean	SD	Min	1st Q	Median	3rd Q	Max
January	201.16	99.71	17	118	211	305	355
February	202.97	103.96	4	117.5	210	294.5	365
March	225.81	95.83	29	166	256	300	362
April	203.67	109.37	2	88.25	225	282.75	351
May	207.97	114.97	31	103	226	313	364
June	195.73	117.87	20	82	207.5	309.5	366
July	181.55	109.61	13	88	188	284	350
August	173.45	112.73	11	61	145	291	352
September	157.30	87.15	1	79.25	168	232.25	315
October	182.45	96.78	5	94	201	244	359
November	148.73	94.40	9	73.5	131.5	209.75	348
December	121.55	95.06	3	43	100	163	328

[1] Adapted from McClave & Benson, case study 3.2.

Even though the process itself seems to be fair, there is a general downward trend in monthly mean and median selection orders (thus disproportionate number of men born later in the year were selected early in the draft). To see this clearly, consider the scatter plot below of the mean rank by month.



If the lottery were truly random than one would expect that the monthly means would be fairly close to the average value of the selection order =
(1 + 2 + 3 + 4 + … +365+366)/366 = 183.5
and they would show no pattern from month to month.
Why did method that seemed so fair produce such a bias? The problem was with the bowl. The January capsules had all been put in first, the February capsules were added next, and so on. The capsules were then mixed but not thoroughly enough. There was still a tendency for dates early in the year to be near the bottom of the bowl. Compounding the problem was that the bowl's compactness made it physically difficult to reach in and select a capsule from beneath the topmost layers. As a result capsules were more or less chosen from the top down. Since composition of the layers was not random because of the earlier inadequate mixing, the final sequence contained the bias evident in the picture above.

Why Randomize?

Results from polls and other statistical studies reported in newspapers or magazines often emphasize that the samples were randomly selected. Why the emphasis on randomization? Couldn't a good investigator do better by carefully choosing respondents to a poll so that various interest groups were represented? Perhaps, but samples selected without objective randomization tend to favor one part of the population over another. For example, polls conducted by sports writers tend to favor the opinions of sports fans. This leaning toward one side of an issue is called *sampling* or *selection bias*. In the long run, random samples seem to do a good job of producing samples that fairly represent the population. In other words, randomization *reduces* sampling bias. Note that **random sampling can be tough.** A random sample is *not* a casual or haphazard sample. The target population must be carefully identified, and an appropriate sampling frame must be selected.

SELECTION BIAS A systematic tendency to over-represent some segment of the population. (Some subject or group of subjects is more likely to be selected by the sampling method than another.)

Example 5.2. Literary Digest 1936

In 1936, the *Literary Digest* mailed 10 million questionnaires to U.S. voters, asking them whether they preferred Franklin Delano Roosevelt or Alf Landon. The *Digest* had been predicting elections since 1916 and boasted of their accurate projections. Received 2.4 million responses -- a sample size approximately 800 times larger than the Gallup Poll.

Sample Results: Landon 57% FDR 43% Actual Election: FDR 62% Landon 38%

What happened?

- Bad sampling frame -- *Literary Digest* subscribers were not representative of the American voting population.
- Nonresponse bias -- 76% of subscribers who didn't return the survey were much more likely to support FDR.
- Also, Landon's supporters, desiring change, tended to be more vocal in their support of their candidate than did FDR's supporters.

Difficulties and Disasters in Sampling (Utts, p. 62)

Difficulties:

1. Using the wrong sampling frame
2. Not reaching the individuals selected
3. Getting no response or getting a volunteer response

Disasters:

1. Getting a volunteer sample
2. Using a convenience or haphazard sample

Example 5.3. The New Hite Report -- Controversy over the Numbers

[Reference: Streitfeld, D. (1988). "Shere Hite and the trouble with numbers." *Chance, 1,* 26-31.]

In 1968, researcher Shere Hite shocked conservative America with her now-famous "Hite Report" on the permissive sexual attitudes of American men and women. Twenty years later, Hite was surrounded by controversy again with her book, *Women and Love: A Cultural Revolution in Progress* (Knopf Press, 1988). In this new Hite report, she reveals some startling statistics describing how women feel about contemporary relationships:

- 84% of women are not emotionally satisfied with their relationship
- 95% of women report "emotional and psychological harassment" from their men
- 70% of women married 5 years or more are having extramarital affairs
- Only 13% of women married more than 2 years are "in love"

Hite conducted the survey by mailing out 100,000 questionnaires to women across the country over a 7-year period. Each questionnaire consisted of 127 open-ended questions, many with numerous subquestions and follow-ups. Hite's instructions read: "It is not necessary to answer every question! Feel free to skip around and answer those questions you choose." Approximately 4,500 completed questionnaires were returned for a response rate of 4.5%, and they form the data set from which these percentages were determined. Hite claims that these 4,500 women are a representative sample of all women in the United States, and therefore, the survey results imply that vast numbers of women are "suffering a lot of pain in their love relationships with men." Many people disagree, however, saying that only unhappy women are likely to take the time to answer Hite's 127 essay questions, and thus her sample is representative only of the discontented.

The views of several statisticians and expert survey researchers on the validity of Hite's "numbers" were presented in an article in *Chance* magazine (Summer 1988). A few of the more critical comments follow.

- Hite used a combination of haphazard sampling and volunteer respondents to collect her [data]. First, Hite sent questionnaires to a wide variety of organizations and asked them to circulate the questionnaires to their members. She mentions that they included church groups, women's voting and political groups, women's rights organizations and counseling and walk-in centers for women. These groups seem not representative of women in general; there is an over-representation of feminist groups and of women in troubled circumstances.... The use of groups to distribute the questionnaires meant that gatekeepers had the power of assuring a zero response rate by not distributing the questionnaire, or conversely of greatly stimulating returns by endorsing the study in some fashion. Second, Hite also relied on volunteer respondents who wrote in for copies of the questionnaire. These volunteers seem to have been recruited from readers of her past books and those who saw interviews on television and

in the press. This type of volunteer respondent is the exact opposite of the randomly selected respondent utilized in standard survey research and even more potentially unrepresentative than the group samples cited above. (**Tom Smith, National Opinion Research Center**)

- So few people responded, it's not representative of any group, except the odd group who agrees to respond. Hite has no assurance that even her claimed 4.5% response rate is correct. How do we know how many people passed their hands over these questionnaires. You don't want to fill it out, you give it to your sister, she gives it to a friend. You'll get one response, but that questionnaire may have been turned down by five people. (**Donald Rubin, Harvard University**)
- When you get instructions to answer only those questions you wish to, you're likely to skip some. Isn't it more likely that, for example, a woman who feels strongly about affairs would more likely answer questions on that subject than a woman who does not feel as strongly? Thus, her finding that 70% of all women married over five years are having affairs is meaningless because she does not report how many people answered each question. I cannot tell whether this means 70% of 1000 women or 70% of 10 women. (**Judith Tanur, State University of New York at Stony Brook**)
- Even in good samples, where you have a 50% or 70% response rate, you usually have some skews -- say, with income, race, or region. If she can do a sample like this, she's got the Rosetta Stone, and I'll come study with her. (**Martin Frankel, Baruch College**, commenting on Hite's claim that her sample matches that of the U.S. female population in terms of demographic balance.)
- According to Hite, whether you're 18 or 71, you're going to answer the questions the same way. Whether white, black, Hispanic, Middle Eastern, or Asian American, you're going to answer the same way. Whether you make $5,000 a year or over $75,000, you'll answer the questions the same way. I've never seen anything like this in my career -- and the Kinsey Institute collects data from everybody. (**June Reinisch, Kinsey Institute, Indiana University**, commenting on Hite's numbers showing that no matter what the demographic breakdown of the women married 5 years or more, about 70% are having extramarital affairs.)

Example 5.4. A Meaningless Poll (Utts, pp. 64-65)

On February 18th, 1993, shortly after Bill Clinton became President of the United States, a television station in Sacramento, California asked viewers to respond to the question: "Do you support the President's economic plan?" The next day the results of a properly conducted study asking the same question were published in the newspaper.

	Television Poll	*Survey*
Yes	42%	75%
No	58%	18%
Not sure	0%	7%

As you can see, those who responded to the television poll were more likely to be those who were dissatisfied with the President's plan. Trying to extend those results to the general

population is misleading. It is irresponsible to publicize such studies, especially without a warning that they result from an unscientific survey and are not representative of general public opinion. You should never interpret such polls as anything other than a count of who bothered to go to the telephone and call.

Haphazard/Convenience Sampling

A few years ago, the student newspaper at a California university announced as a front page headline: "Students ignorant, survey says." The article explained that a "random survey" indicated that American students were less aware of current events than international students were. However, the article quoted the undergraduate researchers, who were international students themselves, as saying that "the students were randomly sampled on the quad." The quad is an open air area where students relax, play frisbee, eat lunch, and so on. There is simply no proper way to collect a random sample of students by selecting them in an area like that. In such situations, the researchers are likely to approach people whom they think will support the results they intended for their survey. Or, they are likely to approach friendly looking people who look like they will cooperate. This is called a *haphazard* sample, and it cannot be expected to be representative.

Nonprobability Sampling

- *Judgment Sample*: Asking various newspaper food critics about which restaurants they would recommend
- *Quota Sample*: If a school has 30% off-campus and 70% on-campus students, attempting to get 30 on- and 70 off-campus students when doing a survey of 100 students
- *Chunk Sample*: If I want to know what students think of a textbook quickly, I could use this class to represent all students using the book.
- The only way to make correct statistical inferences from a population to a sample is to use *probability sampling*...

Investigators Led Astray By Nonrandom Samples[2]

Wainer, Palmer and Bradlow [WPB] (1998) presented a series of situations in which the use of nonrandom samples led or could lead to seriously incorrect conclusions.

- **The Most Dangerous Profession** In 1835, H. C. Lombard published a study on the mean longevity of professions, based on death certificates gathered over more than a half century in Geneva. Each certificate contained the deceased's profession, and age of death. Lombard found that the average age of death for the professions ranged from the early 50s

[2] This material is drawn from Wainer, H., Palmer, S. and Bradlow, E. (1998). "A selection of selection anomalies." *Chance*, **11**(2): 3-7.

to the mid 60s. There was one surprise: the most dangerous profession – with the lowest longevity – was "student" with an average age of 20.7!

- **The Twentieth Century Is A Dangerous Time** WPB collected 204 birth and death dates from the Princeton (NJ) Cemetery. When age at death was plotted as a function of birth year, they found that age of death stayed relatively constant until 1920, when longevity begins to decline rapidly. The average age of death decreases from around 70 years of age in the 1900s to as low as 10 in the 1980s. There must be a reason for this anomaly in the data (what WPB call the "Lombard Surprise") but what?

As WPB indicate, "data obtained by nonrandom sampling occurs often in practice…. We address here … data that have been obtained through *self-selection*; that is, inclusion in the sample is determined by the units themselves and not the data gatherer. In such cases, valid inferences require careful thought about the character of the selection process…. Ignoring self-selection can lead to flawed, often ridiculous findings."

Asking Questions

Suppose we want an honest reading on the number of American citizens who cheat on their taxes? How do we ask people if they cheat?

How you ask the question can seriously affect the answer…

1. Do you agree that unions can cause inconvenience and bad labor-management relations?
2. Do you agree that unions have been important in securing employee rights and decent pay?

Sampling Distributions

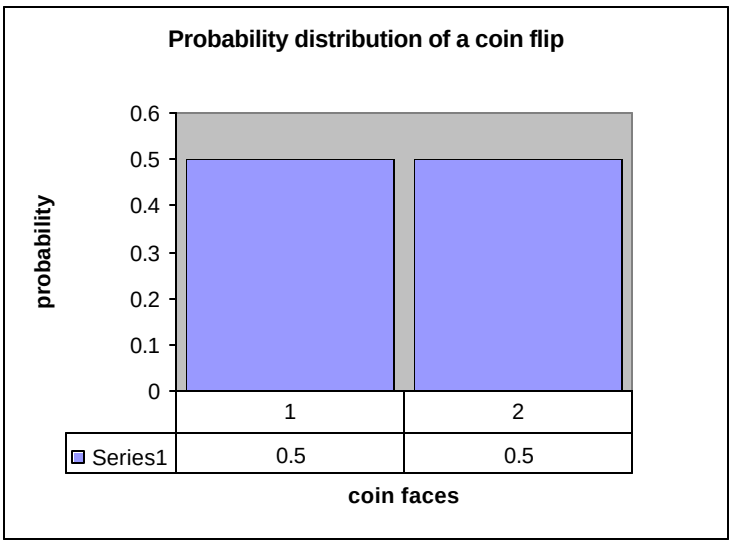
A sample statistic is a random variable, and thus must be compared and judged on the basis of its probability distribution. The probability distribution of a sample statistic is called its *sampling distribution*.

The sampling distribution of a sample statistic (based on n observations) is the relative frequency distribution of the values of the statistic theoretically generated by taking repeated random samples of size n and computing the value of the statistic for each. To illustrate the difference between distribution of some random variable and sampling distribution of statistic of this random variable lets consider following example.

Example 5.5

Consider population created by tossing a fair coin infinitely many times. Let random variable X take value 1 if coin shows tails and 2 if coin shows heads. The probability distribution of X is

x	1	2
$p(x)$	0.5	0.5



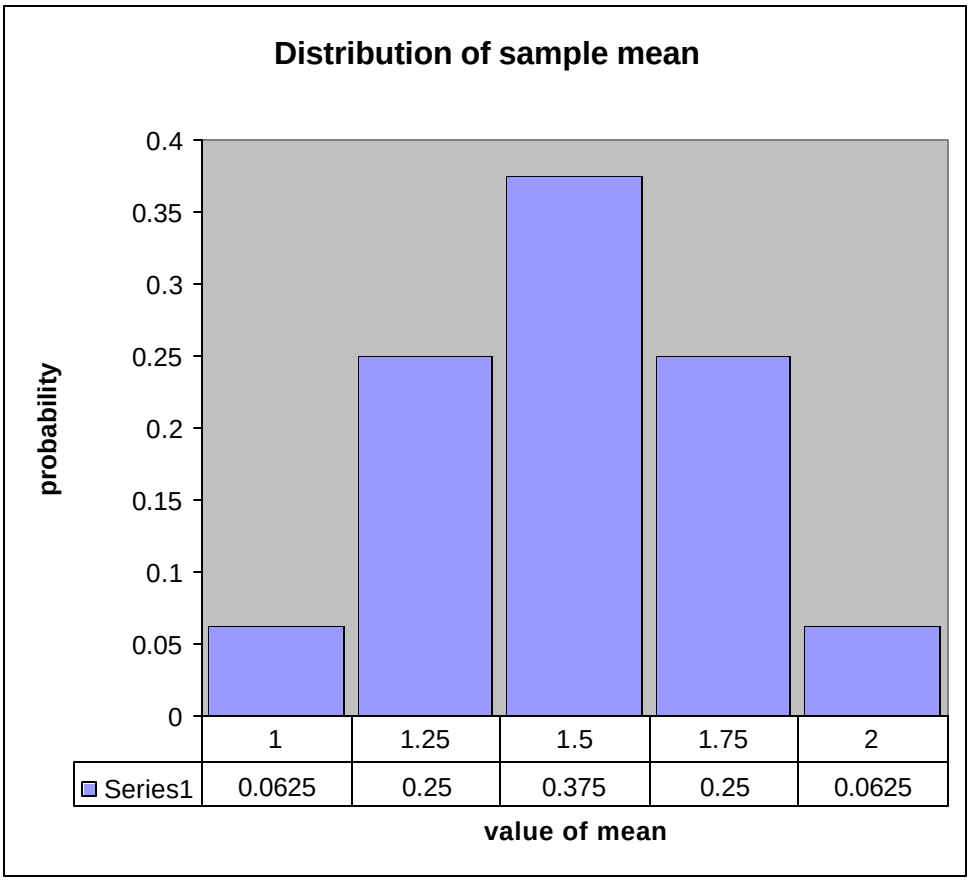
After reading Chapter 3 of this bulkpack you should be able to compute that expected value of X is 1.5 and variance of X is 0.25.

Now suppose that you did not know the mean value of X. You could draw a sample (say of size n = 4) from the population (by flipping coin 4 times) and use sample mean to estimate the population mean. The following table summarizes all possible samples.

Sample	Sample mean
1,1,1,1	1
1,1,1,2	1.25
1,1,2,1	1.25
1,1,2,2	1.5
1,2,1,1	1.25
1,2,1,2	1.5
1,2,2,1	1.5
1,2,2,2	1.75
2,1,1,1	1.25
2,1,1,2	1.5
2,1,2,1	1.5
2,1,2,2	1.75
2,2,1,1	1.5
2,2,1,2	1.75
2,2,2,1	1.75
2,2,2,2	2

Each of these samples has equal chance to come up. Therefore, probability that any one of them occurs is 1/16. Note, however that some of the sample mean values appear more than once. For example, sample mean 1.25 appears 4 times in this table. Hence, it is 4 times more likely to happen compared to sample mean 1 that appears only once. Therefore, probability that sample mean of 1.25 will occur is 4*1/16 =1/4 =0.25. Repeating this argument for other values of sample mean we obtain the following probability distribution of sample mean.

Sample mean	1	1.25	1.5	1.75	2
Probability	0.0625	0.25	0.375	0.25	0.0625



Note that, first of all, sample mean is a random variable itself. Second, distribution of sample mean is very different from the distribution of X. Third, note the bell shape of the sample mean distribution. In fact, the larger the sample size n is the more sample mean distribution will resemble normal distribution (this is the essence of the Central Limit Theorem that we will discuss later in the chapter).
Having the distribution for sample mean we can calculate expected value and variance of the sample mean.

E(sample mean) = 1*0.0625 + 1.25*0.25 + 1.5*0.375 + 1.75*0.25 + 2*0.0625 = 1.5
Var(sample mean) = 0.0625

So, E(sample mean) = $E(X)$, and Var(sample mean) < Var(X). Note that variance of sample mean of X is less than the variance of X itself. This is not a coincidence, in fact, we will see later that sample of size n, Var(sample mean)=Var(X)/n.

The Sampling Distribution of the Sample Mean

Regardless of the shape of the population's distribution:

1. The mean of the sampling distribution of $\overline{X}$ will equal μ, the population mean.
2. The standard deviation of the sampling distribution of $\overline{X}$ is called the *standard error* of the sample mean, and will equal σ, the population standard deviation, divided by the square root of the sample size *n*.

$$\text{Standard Error of the Sample Mean} = s_{\overline{X}} = \frac{s_{pop}}{\sqrt{n}}$$

Example 5.6. Telephone Company[3]

A telephone company knows that during non-holidays the number of calls that pass through the main branch office each hour has a Normal distribution with μ = 80,000 and σ = 35,000. Describe the mean, standard deviation and shape of the sampling distribution of $\overline{X}$, the mean number of incoming calls per hour for a random sample of 60 non-holiday hours.

Solution of example 5.6:

$E(\overline{X}) = \mu = 80{,}000$

$$s_{\overline{X}} = \frac{s_{pop}}{\sqrt{n}} = \frac{35{,}000}{\sqrt{60}} = 4518.481$$

In order to describe shape of the distribution of $\overline{X}$, we will use the following theorem.

Theorem One: Sampling from a Normally Distributed Population

What else might we want to know about the sampling distribution? What can we say about the *shape* of the sampling distribution of the sample mean?

Theorem: If a random sample of *n* observations is taken from a population with a Normal distribution, the sampling distribution of $\overline{X}$ will also be a Normal distribution.

Solution of example 5.6 (continues)

Since population has normal distribution we conclude that $\overline{X}$ is normally distributed too. This along with the values of $E(\overline{X})$ and $s_{\overline{X}}$ that we computed earlier allows us to make the following statement: $\overline{X}$ has normal distribution with mean 80,000 and standard deviation 4518.481.

[3] Adapted from Sincich, exercise 7.24.

Sample Statistics and Sampling Distributions

Probable Error for a Sample Mean

There is no systematic tendency to over- or underestimate μ with $\overline{Y}$. This wouldn't be much of a consolation if we knew that half the time we made a huge overestimate and the other half an equally huge underestimate. The standard error of the sample mean, $s_{\overline{Y}}$, in conjunction with the Empirical Rule, can be used to give a good indication of the probable deviation of a particular sample mean from the population mean.

Example 5.7. (H&O, *BSIFM*, Example 6.3)

Suppose that a supermarket manager is interested in estimating the mean checkout time for the non-express checkout lanes. An assistant manager obtains a random sample of 25 checkout times. If previous data suggest that the population standard deviation is 1.10 minutes, describe the probable deviation of $\overline{Y}$ from the unknown population mean μ.

Solution to Example 5.7. The Empirical Rule indicates that approximately 95% of the time $\overline{Y}$ is within two standard errors $\left(2s_{\overline{Y}}\right)$ of the population mean μ. For $n = 25$,

$$2s_{\overline{Y}} = \frac{2s_{\text{pop}}}{\sqrt{n}} = \frac{2(1.10)}{5} = .44$$

The probable error for $\overline{Y}$ is no more than .44 minute.

The probable accuracy of a sample mean, as measured by its standard error, is affected by the sample size. Because the standard error of the sample mean is the population standard deviation divided by the square root of the sample size, the standard error decreases as the sample size increases. For example, if the sample size had been either 50 or 100 instead of 25 in Example 5.8., the probable errors $\left(2s_{\overline{Y}}\right)$ would have been, respectively, .31 or .22.

Example 5.8. (H&O, Exercise 6.16)

The average demand for rental skis on winter Saturdays at a particular area is 148 pairs, which has been quite stable over time. There is variation due to weather conditions and competing areas; the standard deviation is 21 pairs. The demand distribution seems to be roughly Normal.

a. The rental shop stocks 170 pairs of skis. What is the probability that demand will exceed this supply on any one winter Saturday?

b. The shop manager will change the stock of skis for the next year if the average demand over the 12 winter Saturdays in a season (considered as a random sample) is over 155 or under 135. These limits aren't equidistant from the long-run process mean of 148 because the costs of oversupply and undersupply are different. If the population mean stays at 148, what is the probability that the manager will change the stock?

The Central Limit Theorem: Sampling from Any Population

Basically, the CLT states that if we have a *large enough* sample taken from *any* population, not just a Normally distributed one, then the sampling distribution of the sample mean will also be Normally distributed.

More formally, if a random sample of *n* observations is selected from essentially any population, then, when *n* is sufficiently large, the sampling distribution of $\overline{X}$ will be approximately Normal.

The quality of the Normal approximation to the sampling distribution of $\overline{X}$ depends on two things: the size of the sample, and the shape of the underlying distribution of the population.

- The larger the sample size *n*, the better the Normal approximation to the sampling distribution of $\overline{X}$
- Generally speaking, the greater the skewness of the underlying population, the larger the sample size must be to obtain an adequate Normal approximation.

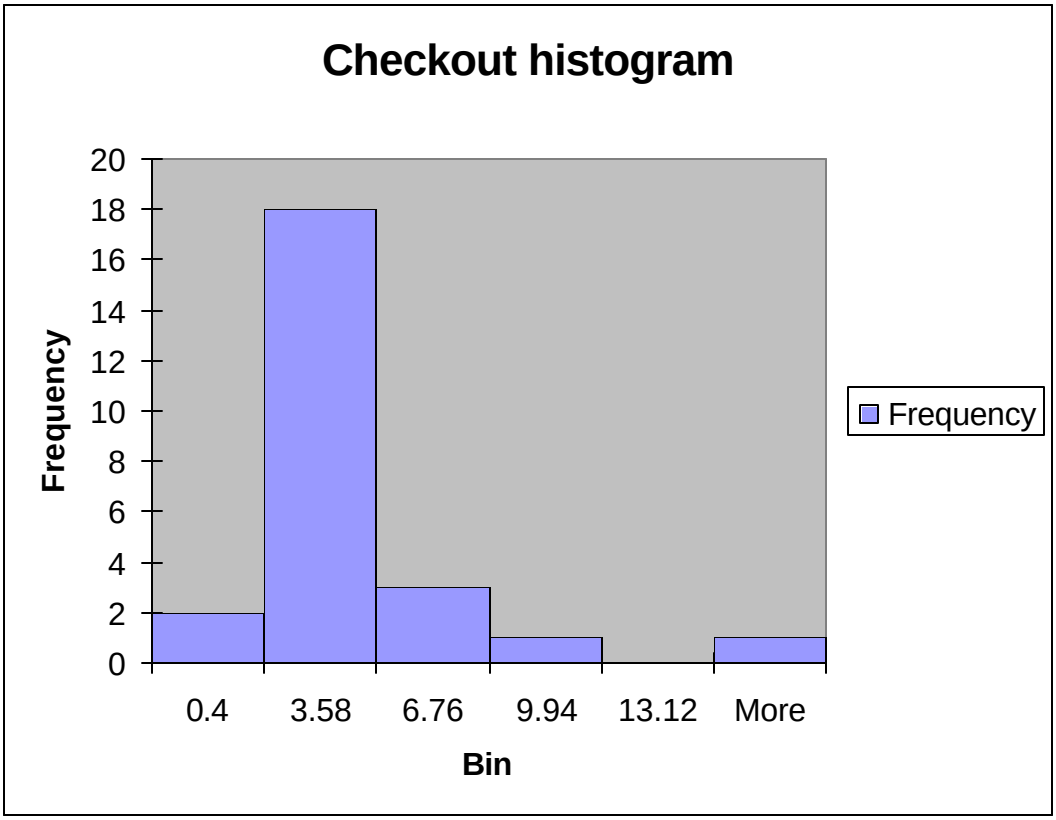
How Large is Large Enough?

A good rule of thumb is based on plots of the sample data.

- If a plot of the sample data shows *severe skewness*, it is reasonable to assume that the underlying population is severely skewed, and the Normal approximation to the sample mean's sampling distribution is not appropriate unless *n* is at least 100.
- For *mild skewness*, $n = 30$ should generally be sufficient to make the Normal approximation to the sampling distribution appropriate.
- For *symmetric but outlier-prone* data, $n = 15$ sample should be enough for the CLT effect to make the Normal approximation reasonable.
- For *normalish* data, $n = 1$ is sufficient.

Example 5.9. (H&O, Example 6.7)

In the supermarket checkout time situation of Example 5.8., the following actual times in minutes were observed ($n = 25$): .4, .4, .5, .5, .5, .6, .6, .7, .8, .9, 1.1, 1.2, 1.4, 1.5, 1.8, 2.0, 2.3, 2.6, 2.9, 3.4, 4.2, 5.0, 6.6, 9.2, 16.3 $\left(\bar{y} = 2.70,\ \text{median} = 1.40, s = 3.56\right)$. Does it appear that a normal approximation to the sampling distribution of $\overline{Y}$ (for future samples of size $n = 25$, for instance) would be satisfactory?



Solution to Example 5.9. The sample data suggest that the population distribution of checkout times is likely to be highly skewed. Most times are quite brief, but there are a few people who really slow things up. (See histogram) A sample of 25 is not enough to *deskew* the sampling distribution. Therefore, the Empirical Rule probabilities (which are based on the Normal approximation) in Example 5.8. are most likely inaccurate for $n = 25$.

Example 5.10. Cents and the CLT – A Lesson worth $8.00

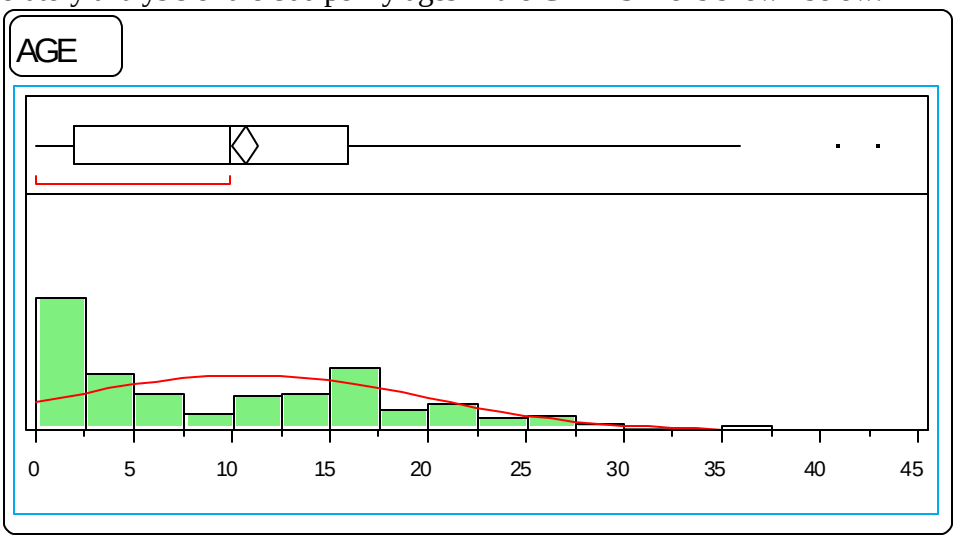
As a consistent example, I collected a sample of 800 pennies in the summer of 1996, and stored the data in the **CENTS** file available on the Course Volume. From the original 800 measurements, samples of size 5, then of size 10, 25 and 40 were drawn (in files **CENTS1**..**CENTS4**), and descriptive statistics calculated:

VARIABLE	N	MEAN	SD	MINIMUM	MAXIMUM	DATA FILE
AGE	800	10.706	9.0771	0.0000	43.000	CENTS
MEAN of 5	160	10.706	4.1451	1.8000	23.800	CENTS1
MEAN of 10	80	10.706	3.0946	4.1000	17.400	CENTS2
MEAN of 25	32	10.706	1.9389	7.4000	14.640	CENTS3
MEAN of 40	20	10.706	1.4406	7.7250	13.100	CENTS4

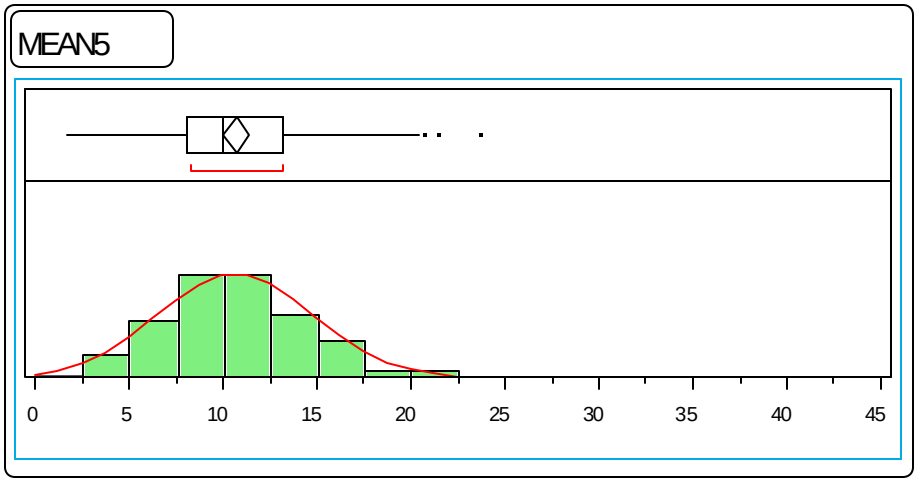
Statistics Pre-Term Program – Chapter Five

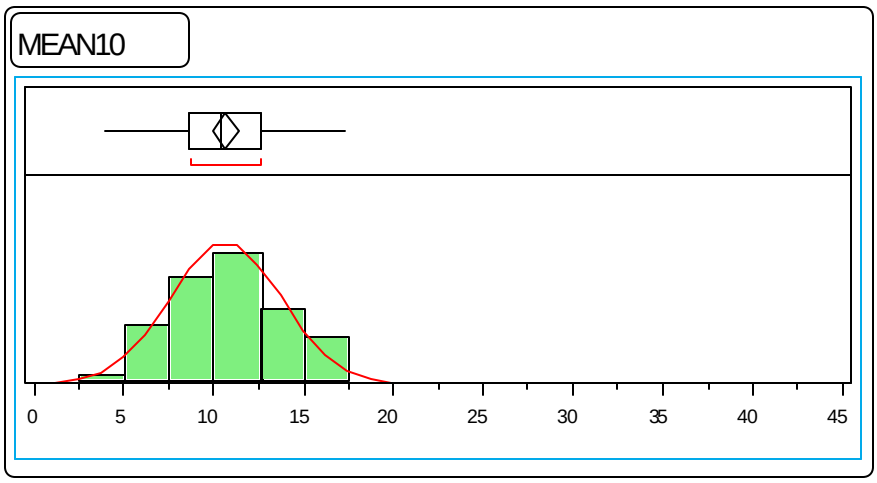
Random Sampling and Sampling Distributions

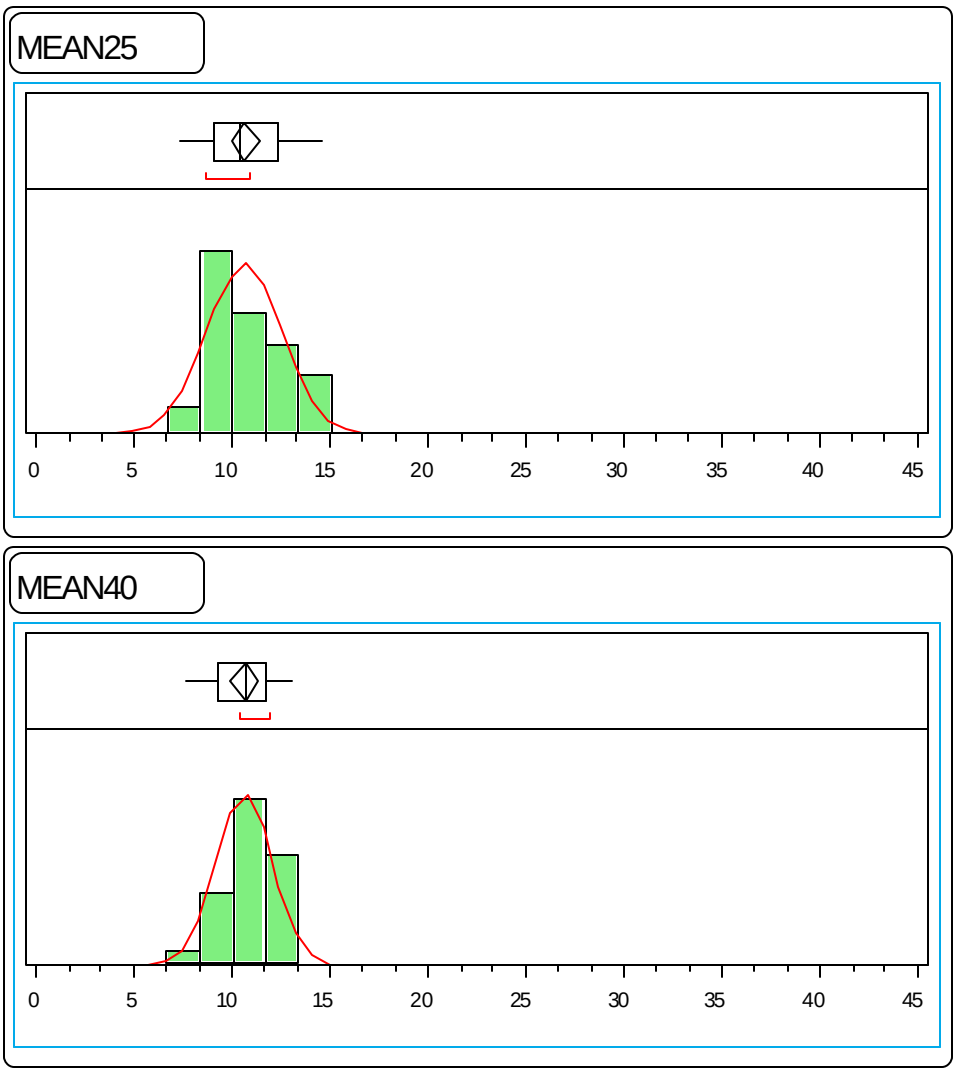
An exploratory analysis of the 800 penny ages in the CENTS file is shown below.



Below are histograms and outlier boxplots of the sample means of sizes 5, 10, 25 and 40. Note that they have been drawn on the same x-axis as the individual ages were above. Note the CLT effect on the *shape*, the *center* and the *spread* of the distribution.




MEAN10
0
5
10
15
20
25
30
35
40
45



Simulation Demonstrations: CLT: How Large is Large Enough?

- We begin with histograms of 1000 means sampled from an exponential population.
  - Note that sample means have a distribution, different than the population. Even for $n = 4$, difference in shape is clear.
  - The average value of the sample means is about 1.0 = population mean for standard exponential.
  - Comparison of scales of the four histograms indicates that variability of means decreases with increasing sample size, but not in proportion to the sample size. (In particular, width of scale for $n = 30$ is not half the width of the scale for $n = 60$).
  - Finally, the CLT effect on the shape of the distribution is obvious.

Next, we summarize (means and standard deviations of sample means) to reinforce ideas.

- The mean of means is a simulation approximation to the expected value. It's approximate because we have 1000 samples, not an infinite number.
- The standard deviation of means is a simulation approximation to the standard error of the mean. The values are quite close to the theoretical values; note that the population standard deviation in this case is 1.

1. Summary Results for Mean of Exponential Sample

Simulation of expected value and standard error (1,000 samples)
Population shape is *skewed*. $\mu = 1.000$, $\sigma = 1.000$
Summary statistics for sampling distribution of the mean:

n	*Mean*	*Std Dev*
4	0.994	0.4747
10	0.998	0.3101
30	1.003	0.1830
60	1.003	0.1278

- Next, we use samples from a **squared exponential** (highly skewed) distribution to demonstrate that the rule $n > 30$ won't work for extreme skew.
- Finally, we use samples from a **Laplace** distribution to show rapid convergence to Normal for outlier-prone, but symmetric populations

2. Summary Results for Mean of Squared Exponential Sample

Simulation of expected value and standard error (1,000 samples)
Population shape is *highly skewed*. $\mu = 2.000$, $\sigma = 4.472$
Summary statistics for sampling distribution of the mean:

n	*Mean*	*Std Dev*
4	2.017	2.2875
10	2.079	1.5147
30	2.037	0.8121
60	2.023	0.5783

3. Summary Results for Mean of Laplace Sample

Simulation of expected value and standard error (1,000 samples)
Population shape is *outlier prone*. $\mu = 0.000$, $\sigma = 1.4142$
Summary statistics for sampling distribution of the mean:

n	*Mean*	*Std Dev*
4	-0.023	0.7029
10	0.010	0.4366
30	0.003	0.2639
60	-0.002	0.1845

Example 5.11. Automobile Batteries

A manufacturer of automobile batteries claims that the distribution of the lifetimes of its best battery has a mean of 54 months and a standard deviation of 6 months. A consumer group purchases a sample of 50 batteries and determines their lifetimes.

a. Assuming the manufacturer's claim is true, describe the sampling distribution of the mean lifetime of a sample of 50 batteries.

A sample of 50 batteries will produce a sampling distribution of the sample mean which follows

- a Normal distribution, (assuming that the population of batteries is not grossly skewed), with
- Expected value $E(\overline{X}) = \mu_{pop} = 54$
- and Standard error of the mean $SE(\overline{X}) = \frac{s_{pop}}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.85$

b. Assuming the manufacturer's claim is true, what is the probability that the group's sample has a mean lifetime of 52 hours or less?

We need to find the probability that the sample mean of 50 batteries, $\overline{X}$, is 52 or less. But, assuming the manufacturer's claim is true, we know that the sample mean $\overline{X}$ follows a Normal distribution with mean $E(\overline{X}) = 54$ and standard error $SE(\overline{X}) = 0.85$. So we have

$$\begin{aligned}\Pr\left(\overline{X} < 52\right) &= \Pr\left(\frac{\overline{X} - E\left[\overline{X}\right]}{SE\left[\overline{X}\right]} < \frac{52 - E\left[\overline{X}\right]}{SE\left[\overline{X}\right]}\right) \\ &= \Pr\left(Z < \frac{52 - 54}{0.85}\right) \\ &= \Pr\left(Z < -2.35\right)\end{aligned}$$

Now Pr(Z < -2.35) can be found using Table 2, and is equal to **.0094**, which is the probability that the group's sample has a mean lifetime of 52 hours or less, assuming the manufacturer's claim is true.

Other Versions of the CLT

Versions of a CLT also apply to sample statistics other than sums and means, but the Normal distribution does not always apply. Mere largeness does not imply Normality, unless a sum or average is involved. The CLT guarantees that the distribution of the *sample mean* will be Normally distributed, even though the individual values may be quite skewed. The best advice is to draw pictures...

Four Questions on Sampling Distributions

1. Suppose we draw a sample of $n = 64$ observations from a new population, with $\mu = 980$ and $\sigma = 300$.
a. Find Pr($\overline{X} > 1{,}050$) b. Find Pr($\overline{X} < 960$) c. Find Pr($\overline{X} > 1{,}100$)

2. An automatic machine in a manufacturing process is operating properly if the lengths of an important subcomponent are Normally distributed, with mean $\mu = 117$ cm and standard deviation $\sigma = 2.1$ cm.
a. Find the probability that a randomly selected unit has a length greater than 120 cm.
b. If 3 units are randomly selected, what is Pr(their mean length exceeds 120 cm.)?

3. The manufacturer of cans of salmon that are supposed to have a net weight of 6 ounces tells you that the net weight is actually a random variable with mean 6.05 ounces and standard deviation of 0.18 ounce. Suppose you take a random sample of 36 cans.
a. Find the probability that the sample mean will be less than 5.97 ounces.
b. Suppose your random sample of 36 cans produces a mean of 5.95 ounces. Comment on the statement made by the manufacturer.

4. The sign on the elevator in a large skyscraper states "Maximum capacity 2,500 pounds, or 16 persons." A professor of statistics wonders what the probability is that 16 people would weight more than 2,500 pounds. If the weights of the people who use the elevator are Normally distributed, with a mean of 150 pounds and a standard deviation of 20 pounds, what is the probability that the professor seeks?

Two Exercises in Sampling Distributions

Example 5.6. Telephone Company (Continued)

Suppose the telephone company wishes to determine whether the true mean number of incoming calls per hour during holidays is the same as for non-holidays. To accomplish this, the company randomly selects 60 hours during a holiday period, monitors the incoming phone calls each hour, and computes $\bar{y}$, the sample mean number of incoming phone calls. If the sample mean is computed to be $\bar{y} = 91{,}970$ calls per hour, do you believe that the true mean for holidays is $\mu = 80{,}000$ (the same as for non-holidays)? Assume that the standard deviation of the number of incoming calls per hour for holidays is 35,000.

Example 5.12. Risk-Taking In Successful Entrepreneurs[4]

A distinguishing characteristic of entrepreneurs is their propensity for taking risks. R. H. Brockhaus used a choice dilemma questionnaire (CDQ) to measure the risk-taking propensities of successful entrepreneurs (*Academy of Management Journal*, September 1980). He found that the CDQ scores of entrepreneurs had a mean of 71 and a standard deviation of 12. (Lower scores are associated with a greater propensity for taking risks.) Let $\bar{y}$ be the mean CDQ score for a random sample of $n = 50$ entrepreneurs.

a. Describe the sampling distribution of $\bar{y}$.
b. Find $\Pr(69 \le \bar{y} \le 72)$.
c. Find $\Pr(\bar{y} \le 67)$. Would you expect to observe a sample mean CDQ score of 67 or lower? Explain.

Answer Sketches for Sampling Distribution Exercises

Telephone Company

To assess whether 80,000 seems reasonable, we need to calculate the probability of obtaining a sample mean of 91,970 or larger when $\mu = 80{,}000$. So we assume that $\mu_{pop} = 80{,}000$ and that $\sigma_{pop} = 35{,}000$ and take a sample of $n = 60$ hours. Again, with a sample of size $n = 60$, we need only to assume that the population of holiday hour incoming calls are not *grossly skewed* in order to allow the CLT effect to occur.

If this is true, then the sampling distribution of $\bar{y}$ is Normal, with Expected value $E(\bar{y}) = \mu_{pop} = 80{,}000$ calls and Standard error $SE(\bar{y}) = \frac{s_{pop}}{\sqrt{n}} = \frac{35{,}000}{\sqrt{60}} = 4{,}518$. Now, we want to find the probability of obtaining a sample mean of 91,970 or larger:

$$\Pr(\bar{y} \ge 91{,}970) = \Pr\left(\frac{\bar{y} - E[\bar{y}]}{SE[\bar{y}]} \ge \frac{91{,}970 - E[\bar{y}]}{SE[\bar{y}]}\right)$$

$$= \Pr\left(Z \ge \frac{91{,}970 - 80{,}000}{4{,}518}\right)$$

$$= \Pr(Z \ge 2.65)$$

and, from Table 2, $\Pr(Z \ge 2.65) = 1 - .9960 =$ **.0040**, which seems awfully small, and thus suggests that the mean number of calls on holiday hours is higher than 80,000.

Risk-Taking In Successful Entrepreneurs

[4] Adapted from Sincich, exercise 7.18.

a. Assuming that the distribution of entrepreneur's CDQ scores in the population is not grossly skewed, the sampling distribution of $\bar{y}$ should be Normal, thanks to the CLT. The expected value of $\bar{y}$ should be equal to the population mean, 71, and the standard error of $\bar{y}$ should be $\frac{12}{\sqrt{50}} = 1.70$.

b.

$$\Pr(69 < \bar{y} < 72) = \Pr\left(\frac{69 - E[\bar{y}]}{SE[\bar{y}]} < \frac{\bar{y} - E[\bar{y}]}{SE[\bar{y}]} < \frac{72 - E[\bar{y}]}{SE[\bar{y}]}\right)$$
$$= \Pr\left(\frac{69 - 71}{1.7} < Z < \frac{72 - 71}{1.7}\right)$$
$$= \Pr(-1.18 < Z < 0.59)$$

and, from Table 2, Pr(-1.18 < Z < 0.59) = .8810 - .2776 = **.6034**

c.

$$\Pr(\bar{y} \le 67) = \Pr\left(\frac{\bar{y} - E[\bar{y}]}{SE[\bar{y}]} \le \frac{67 - E[\bar{y}]}{SE[\bar{y}]}\right)$$
$$= \Pr\left(Z \le \frac{67 - 71}{1.7}\right)$$
$$= \Pr(Z \le -2.35)$$

and, from Table 2, Pr(Z ≤ -2.35) = **.0094**. Thus, we'd expect a sample mean CDQ score of 67 or lower to occur less often than 1 in 100 times. It would be a real surprise to observe such a result.